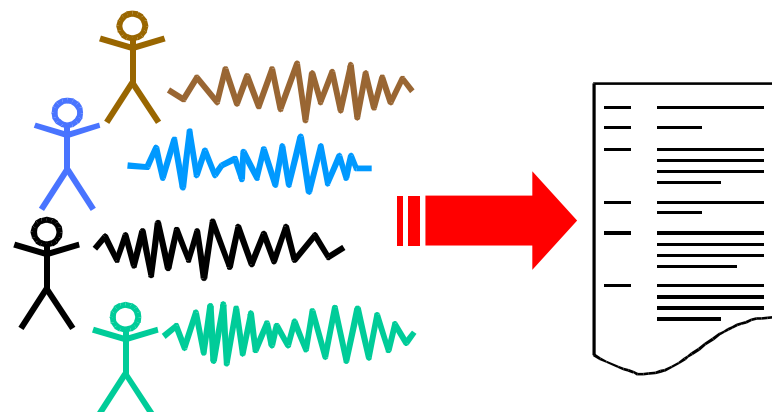


The RT 2003 Fall Metadata Extraction (MDE) Evaluation & Results



Audrey Le, Gregory Sanders, Jonathan Fiscus
NIST Speech Group

RT-03F Workshop
Nov. 13-14, 2003
Washington Marriott

Outline

- Evaluation Overview (Audrey Le)
- Summary of Results
- Analysis of Results (Gregory Sanders)
- Evaluation Tools & Metrics (Jonathan Fiscus)
- Remarks

Evaluation Overview

RT-03 Fall Evaluation

- MDE supplement to the RT-03 Spring STT Evaluation
- Metadata captures information about the speaker, disfluencies in the speaker's speech, and other phenomena
- Metadata enriches STT transcripts to increase usability (i.e. allow them to be rendered to a more readable form)
- Community input resulted in six evaluation tasks
 - Speaker Diarization
 - Speaker **A**tttributed **STT** (SASTT)
 - SU Recognition
 - **SU B**oundary **D**etection (SUBD)
 - Disfluency Recognition
 - **E**dit **W**ord **D**etection (EWD)
 - **F**iller **W**ord **D**etection (FWD)
 - **I**nterruption **P**oint **D**etection (IPD)
 - Composite Task
 - 20**03** **R**ich **T**ranscription (03RT)

Speaker Attributed STT (SASTT)

- Detecting the speaker who spoke each word

Example: 

- Manually created reference

do you think there is some important component of
this that is symbolic I doubt it you know because I
th I doubt they want to just be seen as carrying out
Bill Clinton's policy as the main way in which they
establish themselves

- Manually created reference with speakers segmented

do you think there is some important component of
this that is symbolic I doubt it you know because I th I
doubt they want to just be seen as carrying out Bill Clinton's
policy as the main way in which they establish themselves



SU Structure

- A unit of speech that expresses a complete thought or idea
 - Does not necessarily correspond to a complete sentence
 - Also known as Sentence-like Unit, Syntactic Unit, Semantic Unit, or Slash Unit
- Four SU subtypes
 - Question
 - An SU that functions as a question
 - Example: what about the language barrier
 - Statement
 - An SU that functions as a declarative statement
 - Example: I speak Italian
 - Backchannel
 - An SU that provides acknowledgement to the other speaker
 - Example: okay
 - Incomplete
 - An SU that is incomplete
 - Example: but the thing about (got interrupted by another speaker then resumed with) okay

SU Boundary Detection (SUBD)

- Detecting the boundaries between SU's

Example: 

- Manually created reference with SU boundaries marked by forward slashes (/)

do you think there is some important component of
this that is symbolic / I doubt it / you know
because I th I doubt they want to just be seen as
carrying out Bill Clinton's policy as the main way
in which they establish themselves /

Disfluency Detection Tasks

- Three disfluency detection tasks
 - Edit **W**ord **D**etection (EWD)
 - Filler **W**ord **D**etection (FWD)
 - Interruption **P**oint **D**etection (IPD)


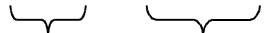
Disfluency Structure

- Portions of speech in a speaker's utterance that are not complete and fluent
- Follows a common structure
 - A **de**letable **p**art **o**f a **d**isfluency (DEPOD)–a portion of the utterance that if deleted does not change the meaning of the utterance
 - An **i**nterruption **p**oint (IP)–the location of a prosodic phenomenon indicating a transition from fluent to non-fluent speech
 - A correction–an optional portion of the utterance that has been corrected
- Two types of disfluencies
 - Edit disfluencies
 - Have a DEPOD, one or more IP's, and optionally a correction
 - Four edit subtypes: repetitions, revisions, restarts, and complex
 - Filler disfluencies
 - Have a DEPOD and an IP
 - Four filler subtypes: filled pauses, discourse markers, explicit editing terms, and asides and parentheticals

Graphical Examples of Disfluencies

- Edit disfluency 

[I] * I speak Italian

DEPOD IP correction

IP is after
the
DEPOD

- Filler disfluency 

* eh what about the language
barrier

IP DEPOD

IP is
before the
DEPOD

Edit Word Detection (EWD)

- Detecting the regions of speech that the speaker repeated, corrected, or abandoned

Example: 

- Manually created reference with edit DEPOD words marked by square brackets ([...])

do you think there is some important component of this
that is symbolic I doubt it you know because [I th]
I doubt they want to just be seen as carrying out Bill
Clinton's policy as the main way in which they
establish themselves

Filler Word Detection (FWD)

- Detecting the regions of speech that contain fillers (i.e., you know, um, uh, etc.)

Example: 

- Manually created reference with filler DEPOD words marked by curly brackets ({ ... })

do you think there is some important component of this that is symbolic I doubt it { you know } because I th I doubt they want to just be seen as carrying out Bill Clinton's policy as the main way in which they establish themselves

Interruption Point Detection (IPD)

- Detecting the locations where fluent speech stops and non-fluent speech starts

Example: 

- Manually created reference with IP's marked by asterisks (*)
do you think there is some important component of this
that is symbolic I doubt it * you know because I th *
I doubt they want to just be seen as carrying out Bill
Clinton's policy as the main way in which they
establish themselves

2003 Rich Transcription (03RT)

- Using all of the previous tasks to produce a rich transcript

Example: 

- Manually created reference with all tasks marked

do you think there is some important component of
this that is symbolic / I doubt it / * { you know }

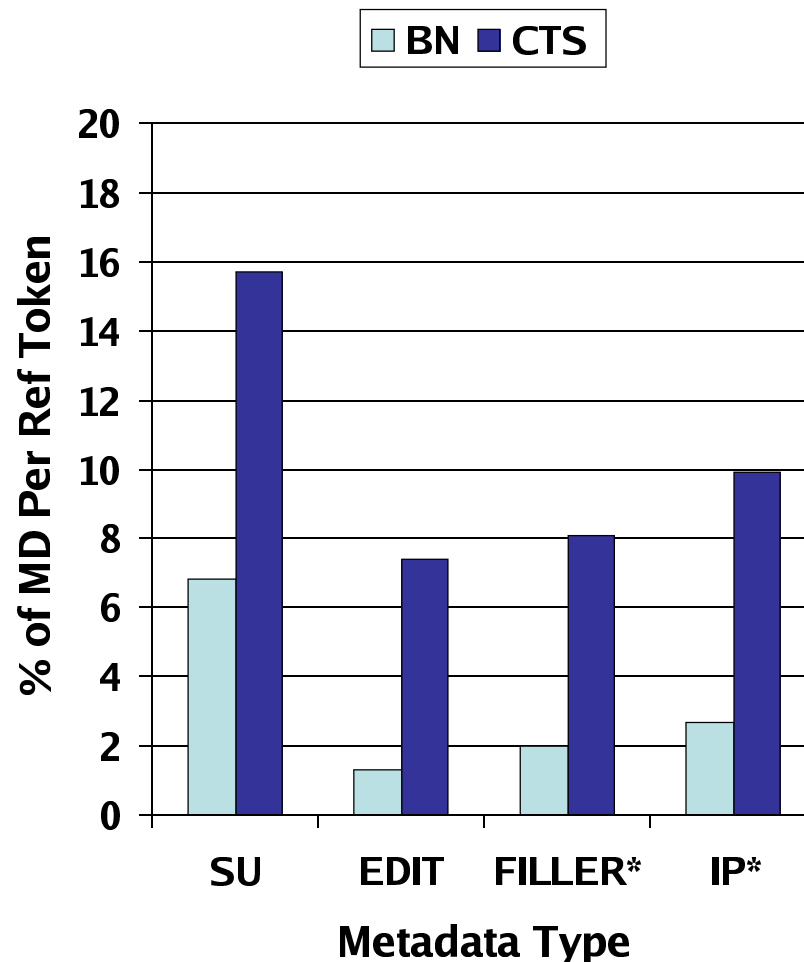
because [I th] * I doubt they want to just be seen as carrying
out Bill Clinton's policy as the main way in which they establish
themselves /

	Speaker A	[...] edit DEPOD words	* IP location
	Speaker B	{ ... } filler DEPOD words	/ SU boundary

Evaluation Corpus

- Half of the English subset of the RT-03 Spring STT Evaluation corpus
- Broadcast News (BN)
 - TDT-4 sources, February 2001 data
 - Three shows, 30-minute excerpt per show
 - 13749 total scorable word tokens (scorable word tokens are LEXEMEs of subtype lexeme, foreign lexeme, fragment, filled pause)
- Conversational Telephone Speech (CTS)
 - Switchboard Cellular and Fisher data
 - 36 conversations, 5-minute excerpt per conversation
 - 35041 total scorable word tokens

Metadata Type by Domain



* indicates rteval and md-eval reported different numbers for FILLER and IP in CTS. Shown are rteval numbers. Equivalent md-eval numbers for FILLER and IP in CTS are 8.07% and 9.83%, respectively.

Evaluation Conditions

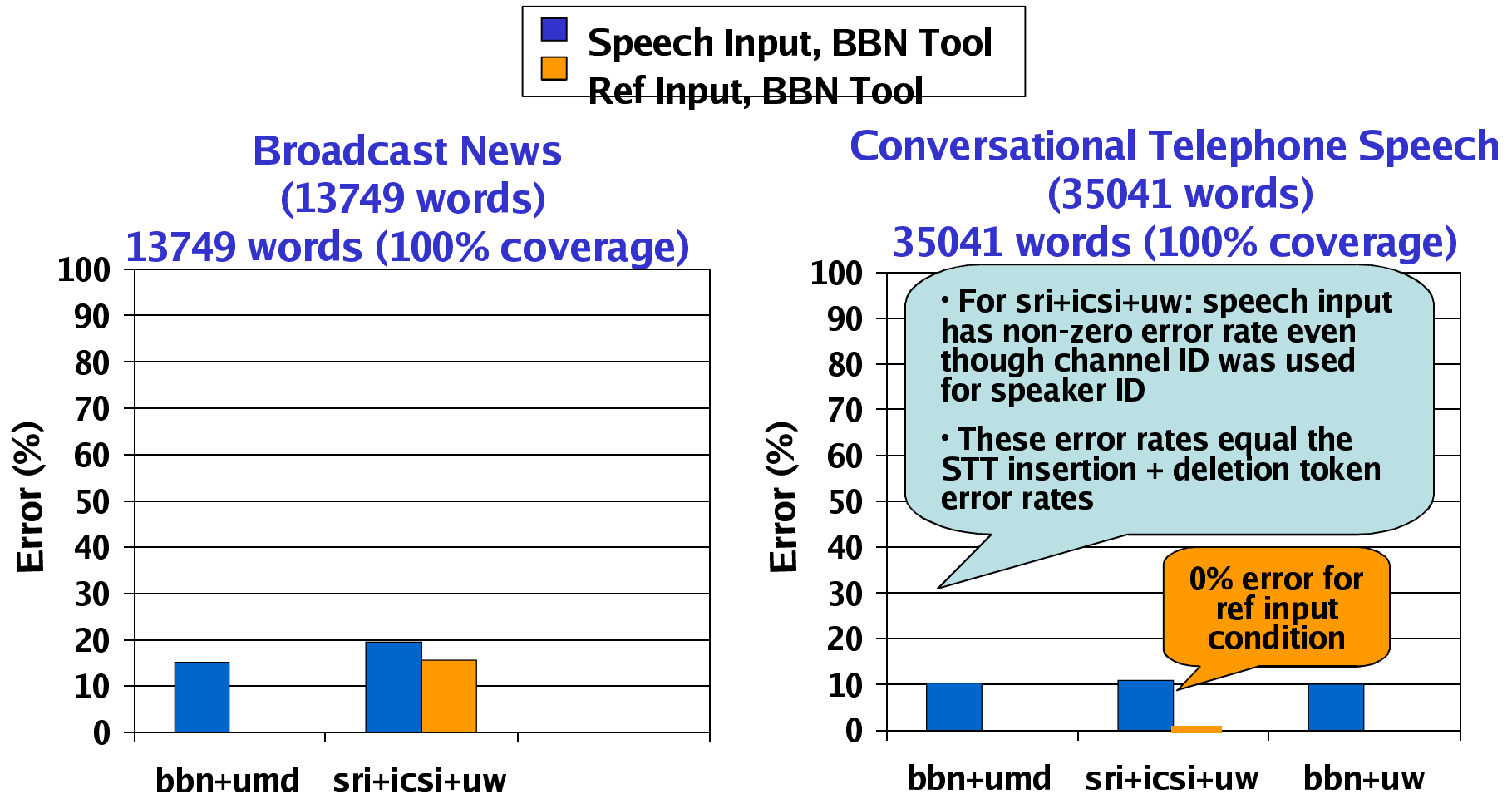
- Speech input
 - Any fully automatic signal processing approach
 - All systems must process the speech input condition
- Speech and reference transcription input
 - Serve as a perfect STT control condition
 - An optional input condition

Participants

- BBN/University of Maryland (bbn+umd)
 - All tasks for speech conditions for both domains
 - Disfluency tasks for reference condition for both domains
- BBN/University of Washington (bbn+uw)
 - All tasks for speech condition for CTS
- Cambridge University (cu)
 - SU boundary detection for speech and reference condition for CTS
- SRI/ICSI/University of Washington (sri+icsi+uw)
 - All tasks for speech and reference conditions for both domains
- *The above indicates primary system submissions only*

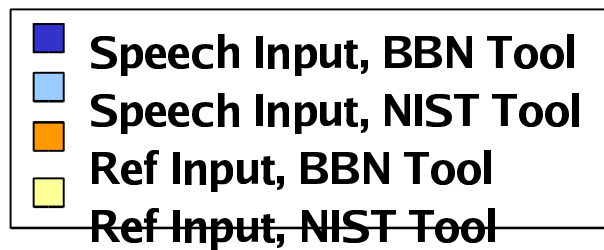
Summary of Results

Speaker Attributed STT Results



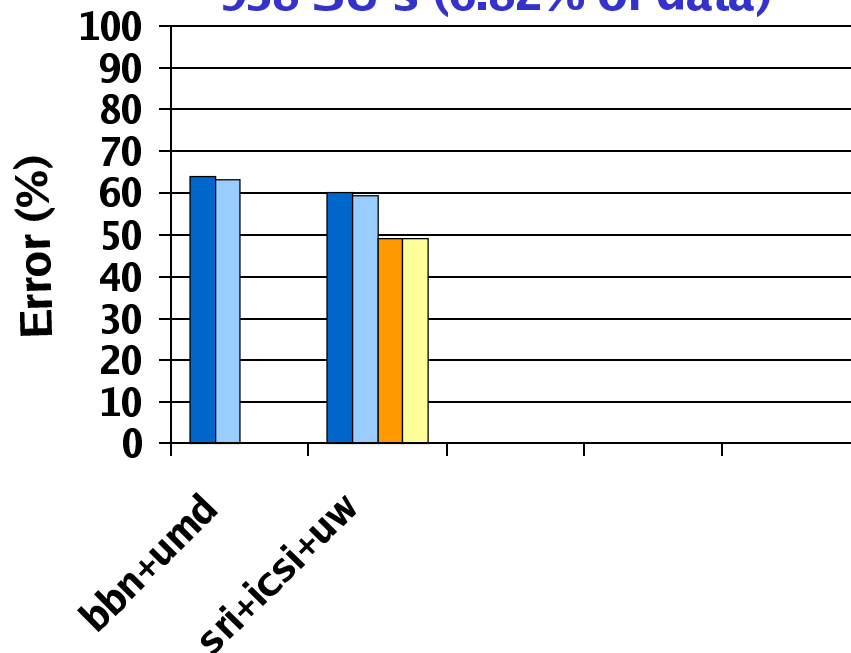
- Should the SASTT error metric measure the speaker diarization or the speaker diarization + STT?
- Current SASTT error metric doesn't do either

SU Boundary Detection Results



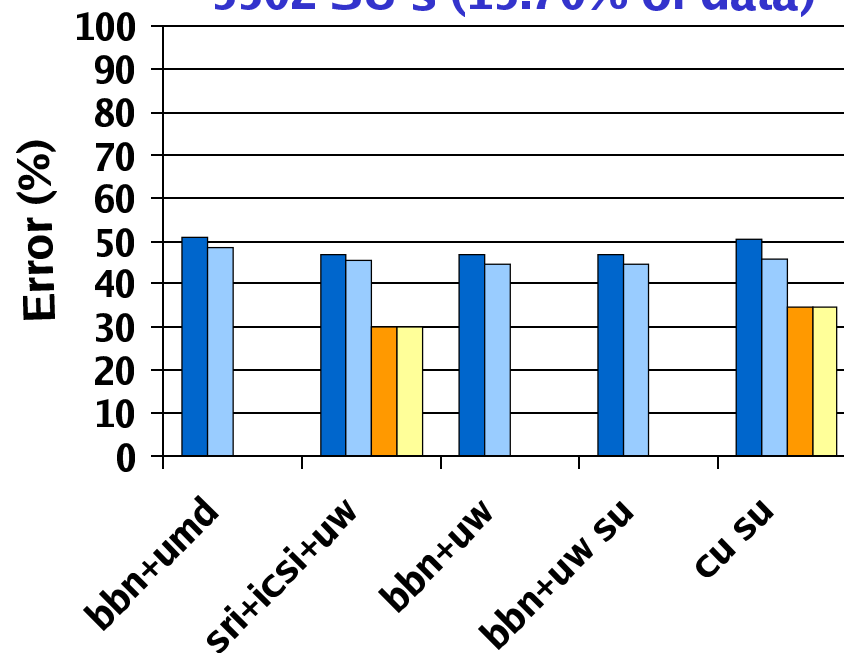
Broadcast News
(13749 words)

938 SU's (6.82% of data)



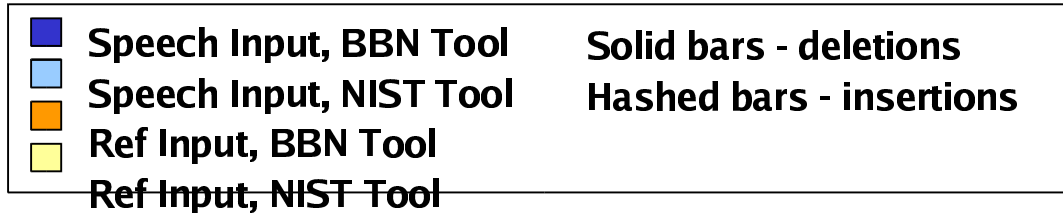
Conversational Telephone Speech
(35041 words)

5502 SU's (15.70% of data)



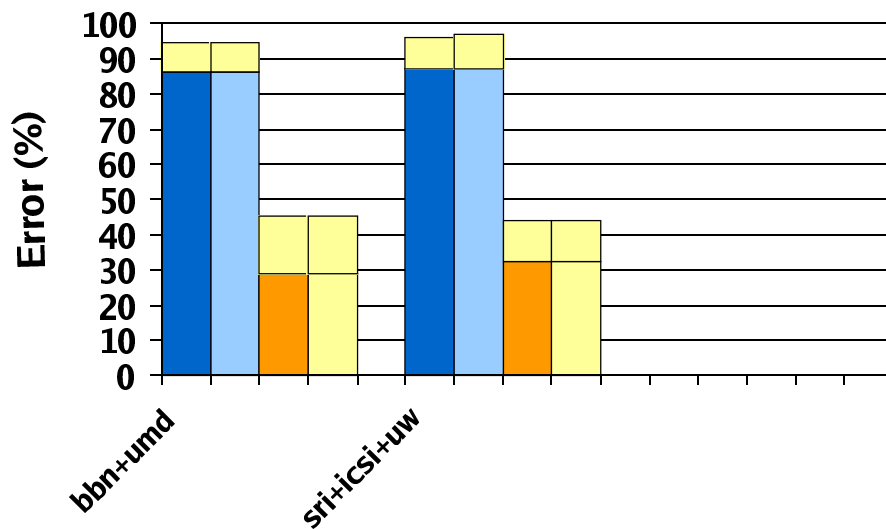
- No essential difference in the error rates reported by the two tools
- Knowing the word tokens helps SU boundary detection (no surprise here)
- SU boundary detection yields lower error rates for CTS

Edit Word Detection Results



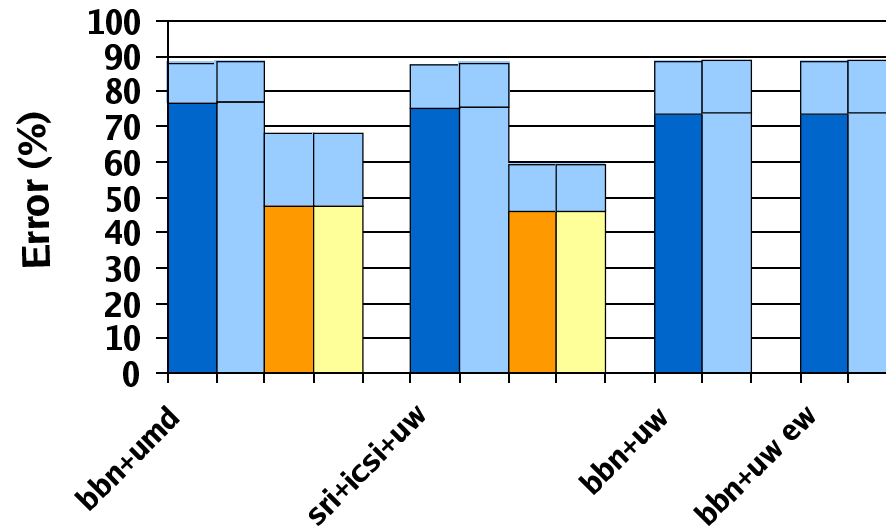
Broadcast News
(13749 words)

181 edit DEPOD words (1.31% of data)



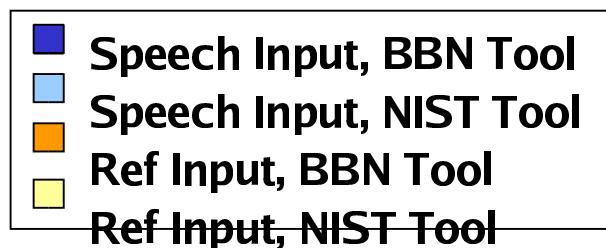
Conversational Telephone Speech
(35041 words)

2587 edit DEPOD words (7.38% of data)



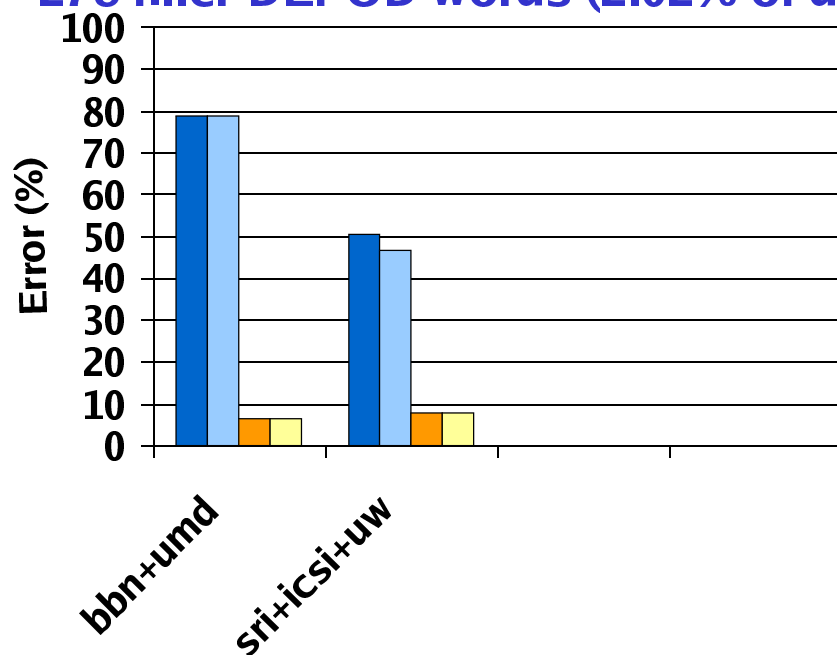
- No essential difference in the error rates reported by the two tools
- Knowing the word tokens helps edit word detection
- Low richness of edit DEPOD's in BN
- Large relative difference between ref versus speech in BN
- Worse performance for speech in BN compared to CTS
- reval: slot deletions of fragments account for most of the errors (>50%)

Filler Word Detection Results



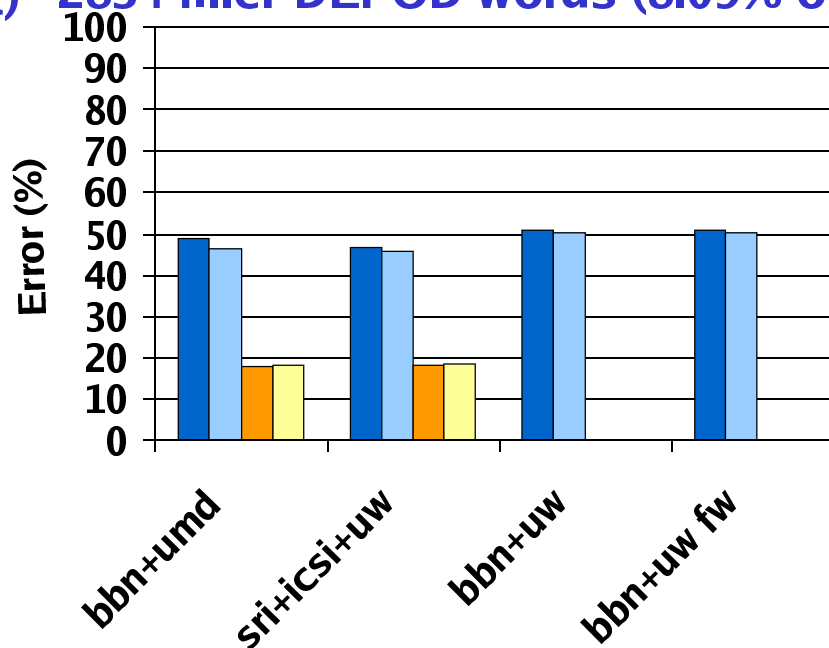
Broadcast News
(13749 words)

278 filler DEPOD words (2.02% of data)



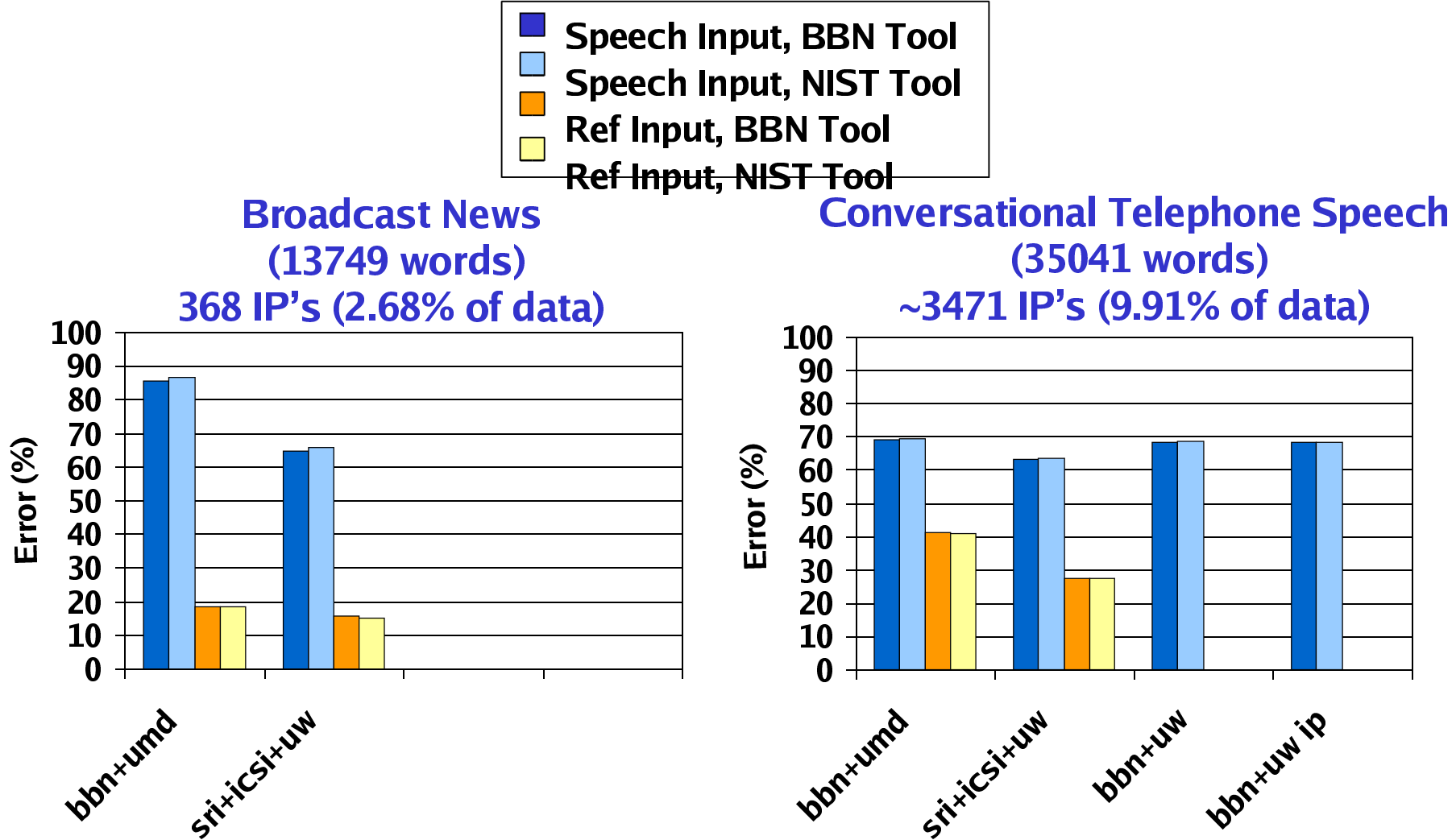
Conversational Telephone Speech
(35041 words)

~2834 filler DEPOD words (8.09% of data)



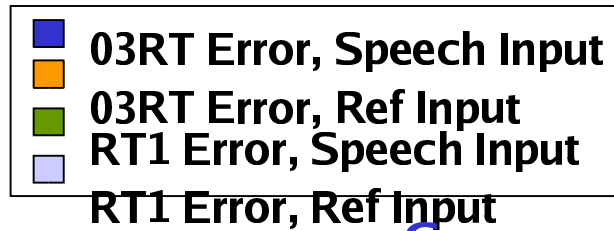
- Large relative difference between ref versus speech in BN
- Very low error rates for ref in BN

Interruption Point Detection Results



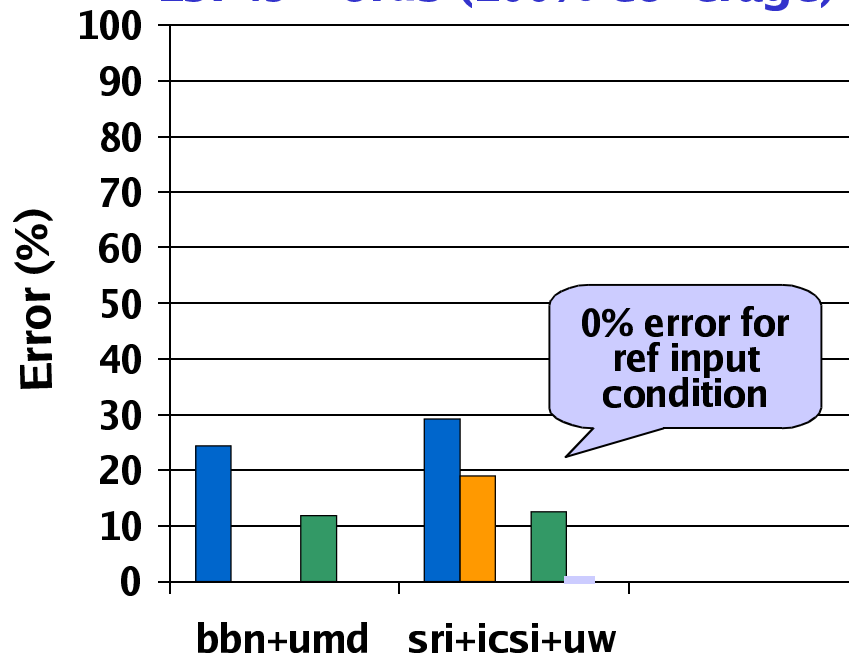
- sri+icsi+uw addressed IP as a separate task (rather than deriving IP's from edits and fillers) and achieved better results

2003 Rich Transcription Results



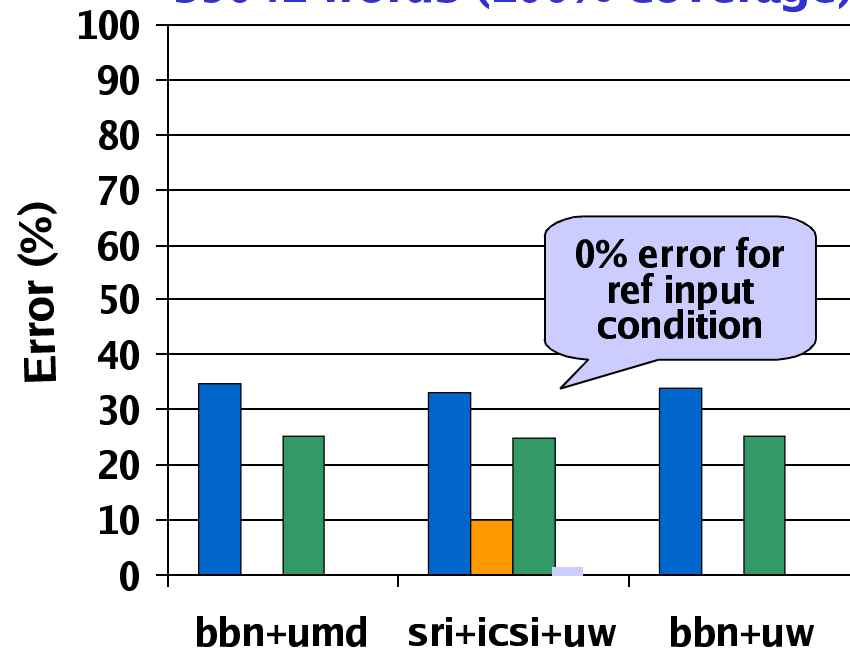
Broadcast News
(13749 words)

13749 words (100% coverage)



Conversational Telephone Speech
(35041 words)

35041 words (100% coverage)



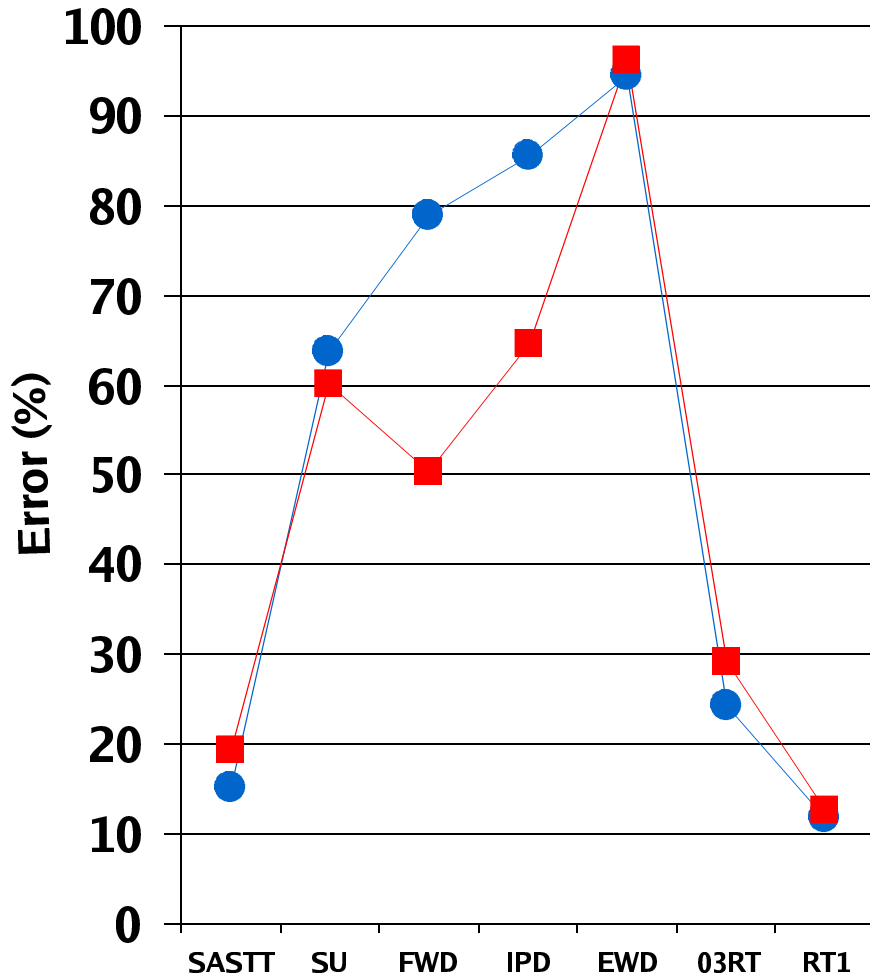
- 03RT error is twice as high as RT1 error for BN

Recap of Results for All Tasks

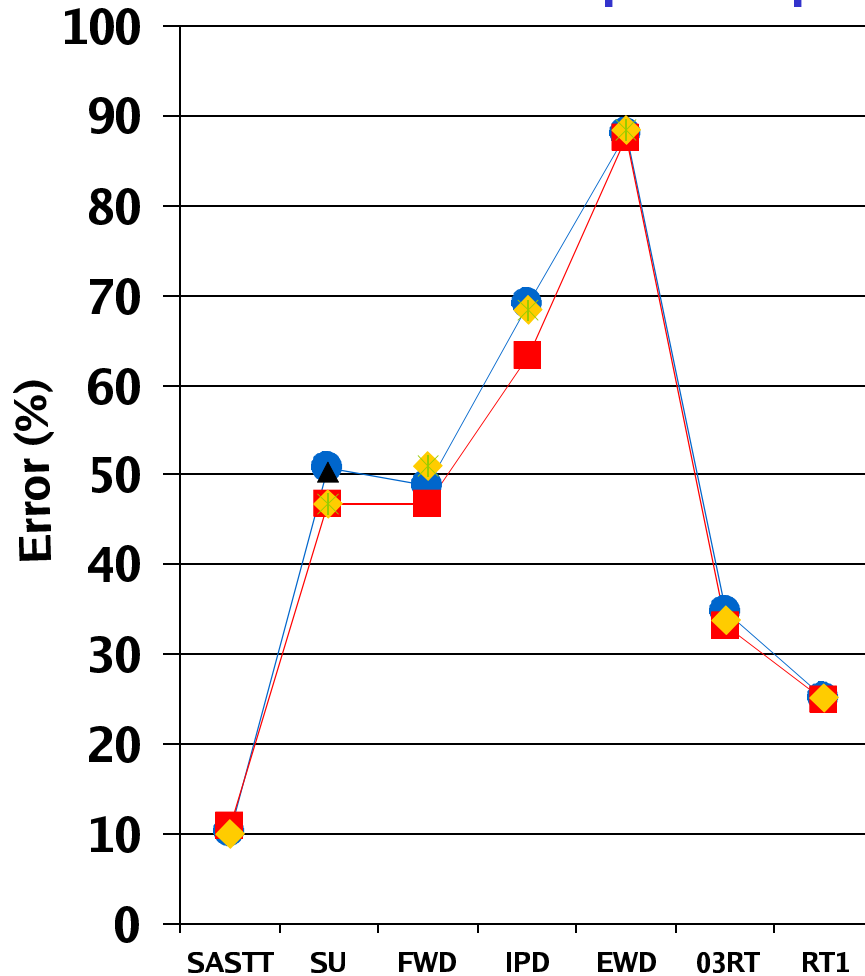
(rteval numbers, speech condition)



Broadcast News

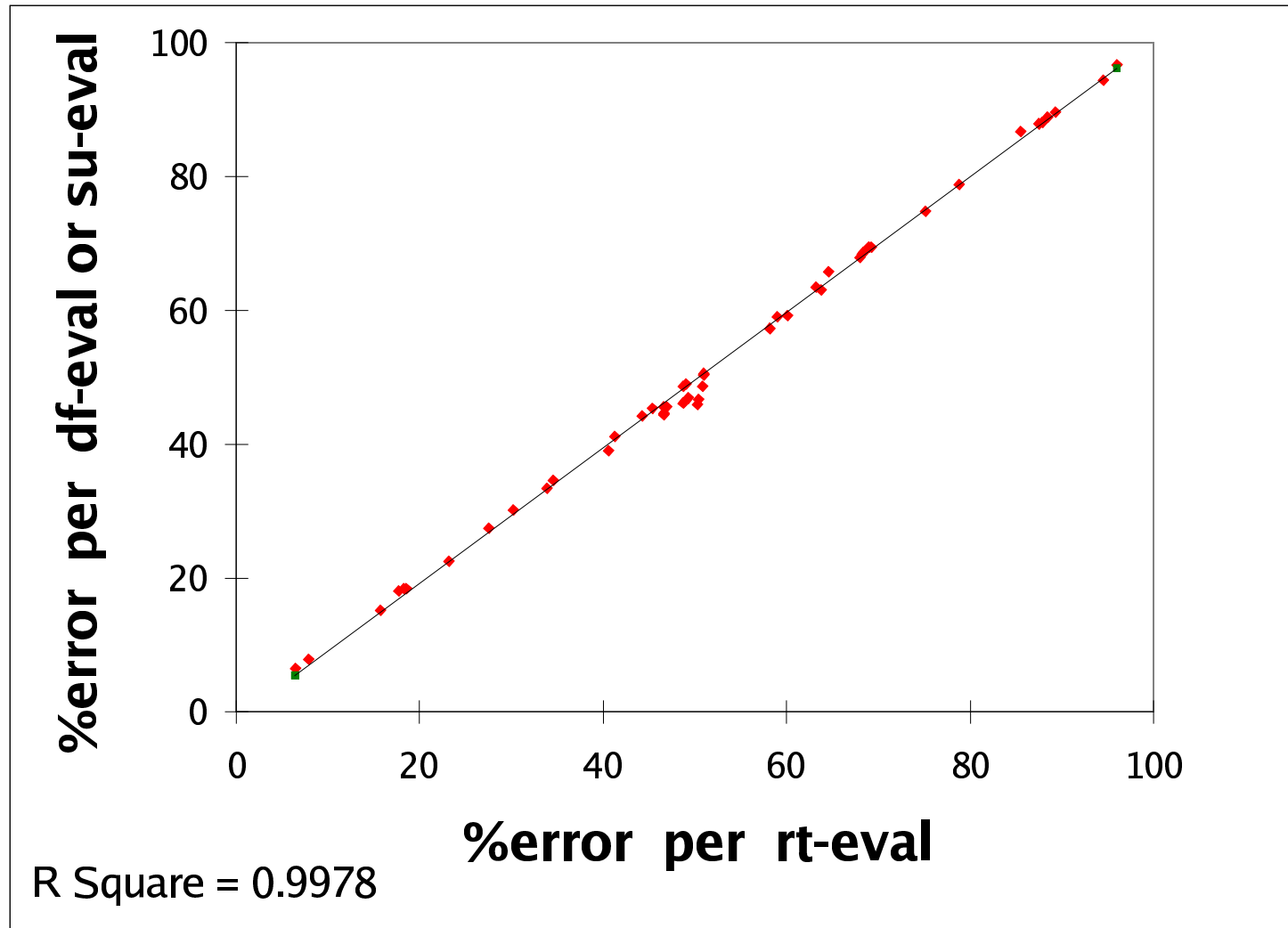


Conversational Telephone Speech



Analysis of Results

Linear Regression Plot for df-eval/su-eval compared to rt-eval



Statistical Significance

- Reporting results of sign test
 - Sign test requires paired samples of error rates (one from each of two systems)
 - Each pair of samples must be independent of the other pairs
 - Minimum requirement is the systems process segments independently
 - Conversational Telephone Speech
 - For each pair of systems, we pair up the error rates on the 72 sides (of the 36 conversations)
 - The sign test shows statistical significance if one system is better than the other on at least 45 of the 72 sides
 - Broadcast News
 - Three Broadcast News shows is not enough samples for the sign test to show significance. We are exploring alternative ways to slice and dice the Broadcast News data.

Speech-To-Text (RT1)

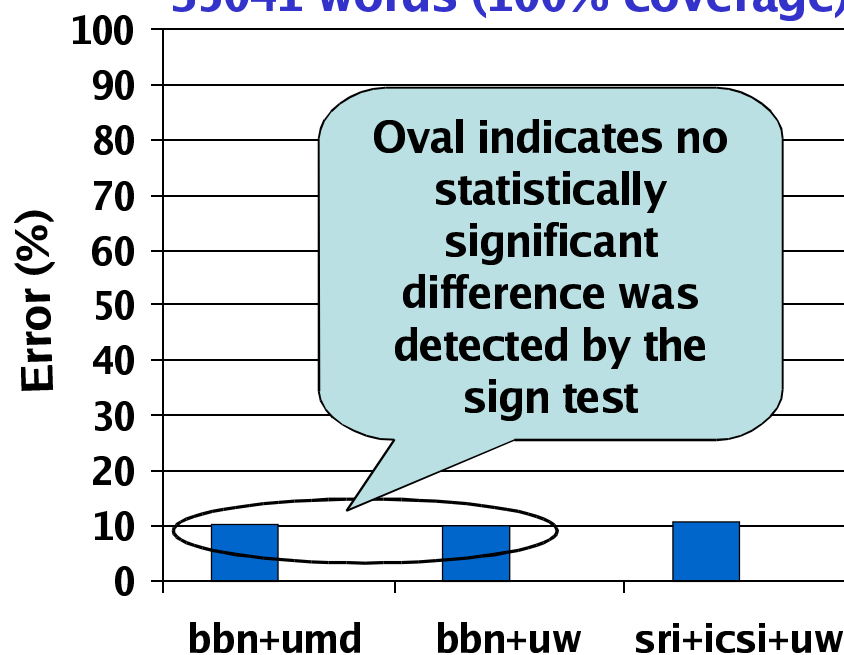
- Sign test does not reveal any statistically significant differences in RT1 performance among the BBN+UMD, BBN+UW, and SRI+ICSI+UW systems on the Conversational Telephone Speech data
 - $p = 0.56$ for SRI+ICSI+UW compared to BBN+UW
 - $p = 0.56$ for SRI+ICSI+UW compared to BBN+UMD
 - $p = 0.29$ for BBN+UW compared to BBN+UMD

Speaker Attributed STT Results

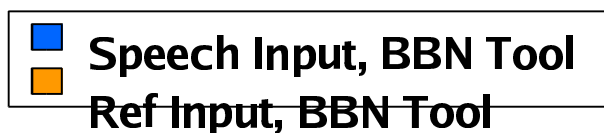


Conversational Telephone Speech
(35041 words)

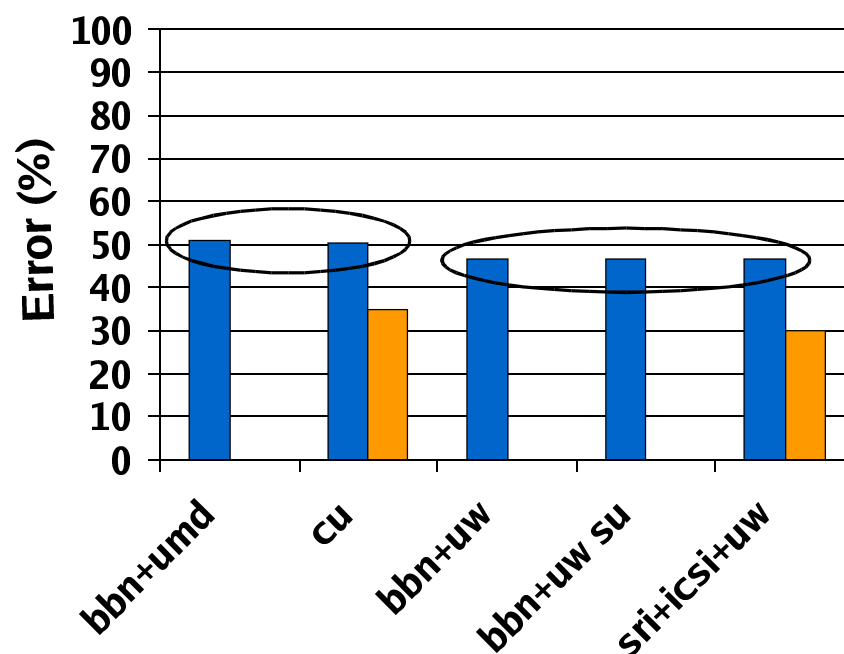
35041 words (100% coverage)



SU Boundary Detection Results



Conversational Telephone Speech
(35041 words)
5502 SU's (15.70% of data)

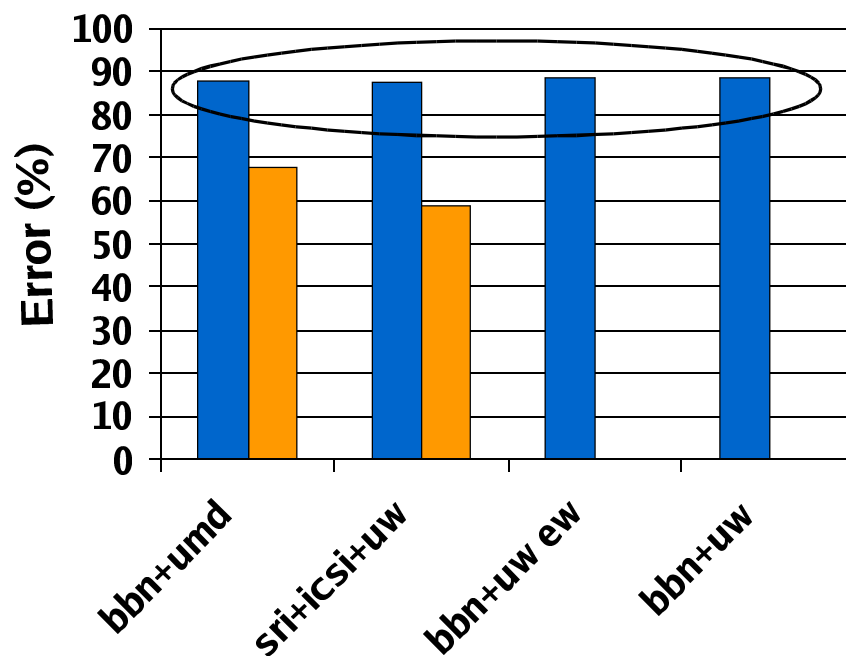


Edit Word Detection Results



Conversational Telephone Speech
(35041 words)

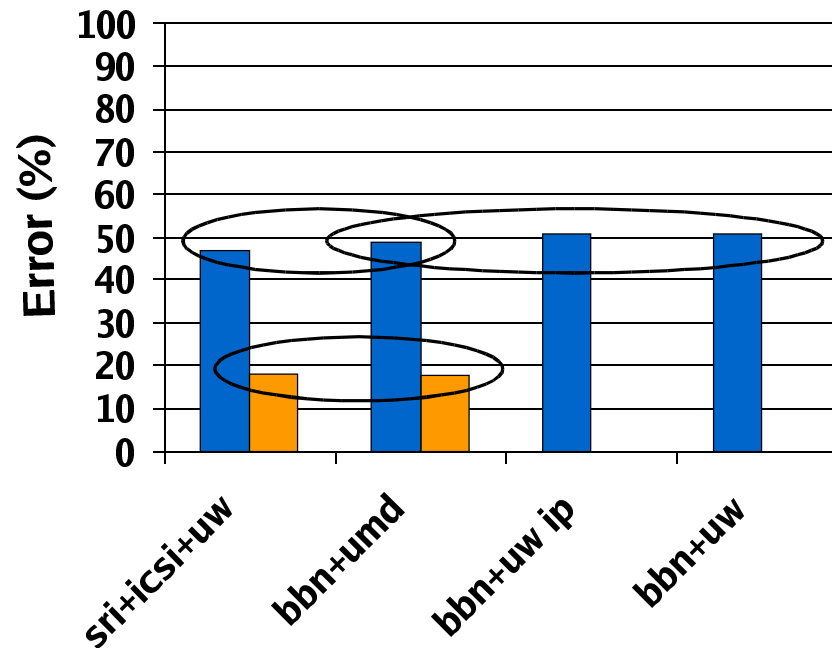
2587 edit DEPOD words (7.38% of data)



Filler Word Detection Results



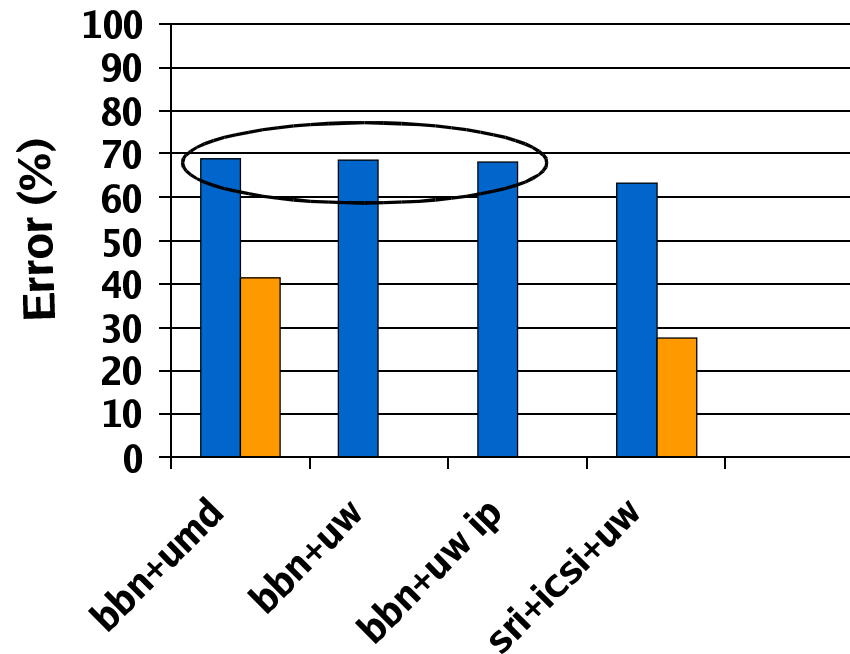
Conversational Telephone Speech
(35041 words)
~2834 filler DEPOD words (8.09% of data)



Interruption Point Detection Results



Conversational Telephone Speech
(35041 words)
~3471 IP's (9.91% of data)

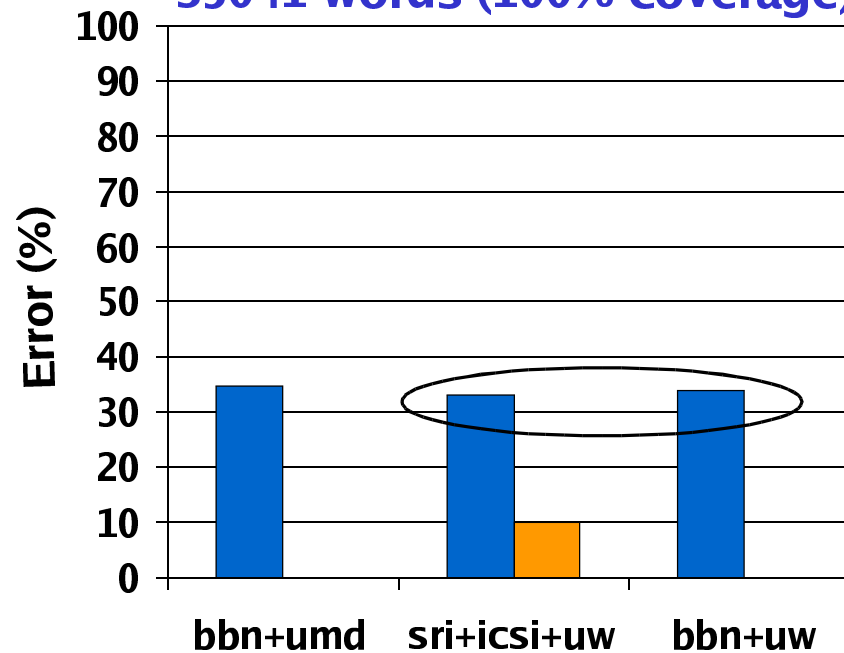


2003 Rich Transcription Results



Conversational Telephone Speech
(35041 words)

35041 words (100% coverage)

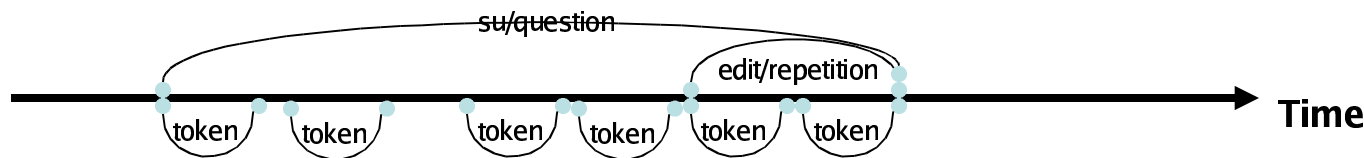


Evaluation Tools & Metrics

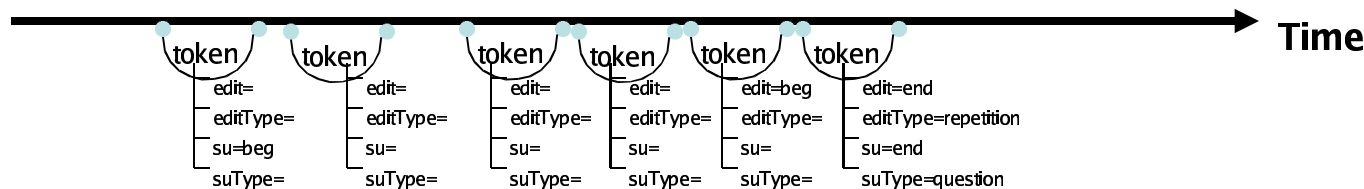
Metadata Representation

(Two Views of the Metadata World)

- Metadata is represented as an object
 - Tokens, SU's, edit DEPOD's etc. are objects
 - Each has a temporal extent and attributes
 - File format supporting this view: RTTM



- Metadata is represented as an attribute on a word
 - Tokens are objects that have temporal extent
 - The attributes (aka 'slots') carry the metadata
 - File format supporting this view: RT-XML



Evaluation Tools

(Two Views of Scoring Metadata)

- Metadata is represented as an object
 - NIST (md-eval)
 - Accepts RTTM format
 - Scores SUBD, EWD, FWD, and IPD tasks
 - Alignment: system output metadata are mapped to the reference metadata
 - Official option: metadata times are adjusted to agree with the aligned system output word tokens
 - Optional option: system output metadata times are used directly
 - Errors are counted in terms of incorrectly identified reference word tokens
- Metadata is represented as an attribute on a word
 - BBN (rteval)
 - Accepts RT XML or RTTM format
 - Scores SASTT, SUBD, EWD, FWD, IPD, and 03RT tasks
 - Two alignments:
 - System output scorable word tokens are aligned to the reference scorable word tokens
 - System speaker labels are mapped to the reference speaker labels
 - Token Error Rate computed for 03RT, Slot Error Rate for the other five
- Both tools were used to score the submissions

Evaluation Task Metrics

- 6 tasks, 10 metrics = 4 from md-eval, 6 from rteval
 - Fully defined in the evaluation plan
- 4 metric types => 2 from md-eval, 2 from rteval

MD-EVAL

Word Coverage Error

- Applies to EWD and FWD

$$\text{Error} = \frac{\# \text{ ref DEPOD tokens not covered by sys DEPODs} + \# \text{ ref non-DEPOD tokens covered by sys DEPODs}}{\# \text{ ref DEPOD tokens}}$$

Boundary Error

- Applies to IPD and SUBD

$$\text{Error} = \frac{\# \text{ missed boundary tokens} + \# \text{ false alarm boundary tokens}}{\# \text{ ref boundary tokens}}$$

- Factors out STT errors

RTEVAL

Slot Error

- Applies to SASTT, SUBD, EWD, FWD and IPD

$$\text{Error} = \frac{\# \text{ sys $TASK tokens that fail to align to ref $TASK tokens} + \# \text{ ref $TASK tokens that fail to align to sys $TASK tokens}}{\# \text{ ref tokens with an active slot}}$$

03RT Token Error

- Applies to 03RT

$$\text{Error} = \frac{\# \text{ inserted system tokens} + \# \text{ deleted reference tokens} + \# \text{ mapped ref/sys tokens with non-matching text or mismatched slots}}{\# \text{ ref tokens}}$$

- Slot Error partially factors out STT errors (i.e., substitutions of token text values do not count)
- RT Token Error fully combines STT and MD Errors

Remarks

Preparations for RT-04

- New tasks?
 - Just say ***NO!***
- Discontinue tasks?
 - Speaker Diarization “Who Spoke When” and IP Detection don’t directly impact readability.
- Updating current tasks?
 - Add SU, Edit and Filler subtype recognition?
 - Does Speaker Attributed STT represent the right task?
The right performance measure?
 - Evaluation plan clarification:
 - Primary vs. non-primary systems
 - Required vs. contrast evaluation conditions

Take Away Messages

- The community successfully ran a metadata evaluation on six tasks
- Metadata error rates for speech input are high
- Disfluency frequency in Broadcast News is low
 - Requires more data for significant results
 - May be insufficient to warrant further research
- SASTT evaluation metric needs to be fixed
 - It should measure either word-based “who spoke when” or the combination of both speaker diarization and STT
- Multiple evaluation tools impeded progress